The BBN RT03 BN Mandarin System

Long Nguyen, Bing Xiang, Dongxin Xu

RT-03 Workshop

Boston, MA, May 19-21, 2003

Overview

- Development test set
- Improvements
- Evaluation results

BBN TECHNOLOGIES
A Verizon Company

Development Test Set

- Selected five episodes from the five audio sources in the TDT4 Mandarin corpus
  - CNR, CTV: radio and TV shows from China
  - CBS, CTS: radio and TV shows from Taiwan
  - VOA: Uncle Sam's radio show
  - Broadcast in the second half of Dec '00
  - First 30 minutes from each episode (~2.5 hours)

	CNR	CTV	VOA	CBS	CTS
Baseline (GI, 1xRT)	11.7	13.0	13.0	57.6	71.8

Quick Fixes

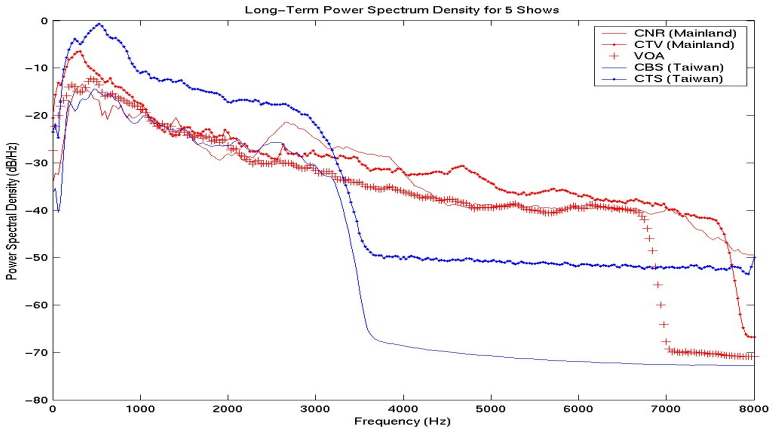
- Used a small phoneme set (73 instead of 160), GD, MLLR adaptation, and ran at 10xRT
- Rebuilt audio segmentation subsystem

	CNR	CTV	VOA	CBS	CTS
0. Baseline	11.7	13.0	13.0	57.6	71.8
1. 73-ph, GD, adapt, 10x	9.4	11.5	10.6	39.4	69.6
2. + rebuilt auto-seg	9.4	11.2	10.2	36.0	63.2

Why are Taiwanese shows (CBS, CTS) so hard?

Data-Specific Issue

- For some unknown reason, the audio for the two Taiwanese shows sound like narrow-band data



BBN TECHNOLOGIES
A Verizon Company

Language-Specific Issue

- Mainland Mandarin is *different* from Taiwanese Mandarin, at least at 3-gram perplexity

CNR	CTV	VOA	CBS	CTS
183	173	269	1265	1265

- Uncle Sam's Mandarin is understandably more Mainland-like

Development solution: build two separate systems

BBN TECHNOLOGIES
A Verizon Company

Beijing System: Language Model

- Training data
  - TDT2, 3, and 4 + People Daily newspaper (91-97, 99, 00) + CNR transcripts (94-96) + Xinhua News text (94-96)
- Lower 3-gram perplexities

CNR	CTV	VOA	CBS	CTS
183	172	269	1265	1265
128	131	222	721	1249

- Lower CERs

	CNR	CTV	VOA
2. + rebuilt auto-seg	9.4	11.2	10.2
3. + Beijing LM	8.2	10.6	9.4

Beijing System: Acoustic Model

- Beijing AM (trained on 93 hours of data)
  - 27 hours H4 corpus
  - 66 hours TDT4 (CNR, CTV, and VOA only) extracted using the same lightly-supervised decoding done for the BN English system

	CNR	CTV	VOA
3. + Beijing LM	8.2	10.6	9.4
4. + Beijing AM	7.8	9.0	7.2

Beijing System: MMI and etc.

- Used the same modern acoustic training as used in the BN English system (SAT, HLDA, MMI)
- Used 4-gram rescoring

	CNR	CTV	VOA	CBS	CTS
4. + Beijing AM	7.8	9.0	7.2	na	na
5. + SAT and MMI	7.3	8.3	7.1	na	na
6. + 4-gram rescoring	7.2	8.1	6.7	na	na

Taipei System

- AM trained on the same 93hrs but band-limited
- LM estimated on Beijing data + Taipei web data (CDN text and CTS transcripts from 1997-2000)
  - Lower 3-gram perplexities (for Taiwanese shows)

CNR	CTV	VOA	CBS	CTS
128	131	222	721	1249
168	147	202	418	276

	CBS	CTS
2. Rebuilt auto-seg	36.0	63.2
7. + narrow-band AM	27.8	57.0
8. + Taipei LM	24.6	46.7

Taipei System: Better AM

- Extracted ~10hrs from TDT4's CBS and CTS shows using lightly-supervised decoding
- Pooled (93h + 10h) and retrained, then MLLR-adapted to the 10hrs of Taiwanese data
  - (old) ML training only
  - No 4-gram rescoring

	CBS	CTS
8. + Taipei LM	24.6	46.7
9. + 10h CBS & CTS	23.2	43.5
10. + MLLR-adapted	22.4	41.6

Roadmap of Improvements

	CNR	CTV	VOA	CBS	CTS
0. Baseline (GI, 1x)	11.7	13.0	13.0	57.6	71.8
1. 73-ph, GD, 10x	9.4	11.5	10.6	39.4	69.6
2. + rebuilt auto-seg	9.4	11.2	10.2	36.0	63.2
3. + Beijing LM	8.2	10.6	9.4		
4. + Beijing AM	7.8	9.0	7.2		
5. + SAT and MMI	7.3	8.3	7.1		
6. + 4-gram rescoring	7.2	8.1	6.7		
7. narrow-band AM				27.8	57.0
8. + Taipei LM				24.6	46.7
9. + 10h CBS & CTS				23.2	43.5
10. + MLLR-adapted				22.4	41.6
Rel. CER reduction	38.5%	37.7%	48.5%	61.1%	42.1%

Benchmark Results

- System ran at 7.5xRT
- Overall CER is skewed by Taiwanese data

	CNR	CTV	VOA	CBS	CTS	All
Dev03 (2.5h)	7.2	8.1	6.7	22.4	41.6	17.7
Eval03 (1h)	3.8	6.8	9.5	21.6	50.6	19.1

Summary

- Achieved significant CER reduction for the Development test set (30% - 60% relative)
- Handled dialect-specific and data-specific issues effectively